

This report is based on research conducted by the XXX Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. <XXX>). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. No statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

This report may be used, in whole or in part, as the basis for the development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products or actions may not be stated or implied.

Methods Research Report

Number xx

Validity and inter-rater reliability testing of quality assessment instruments

Prepared for:

Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No.:

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted for which further reproduction is prohibited without the specific permission of copyright holders.

None of the investigators have any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Author I, Author I. Comparative Effectiveness of Title. Comparative Effectiveness Review No. xx. (Prepared by the <Name> Evidence-based Practice Center under Contract No. xxx-xx-xxxx.) AHRQ Publication No. xx-EHCxxx. Rockville, MD: Agency for Healthcare Research and Quality. <Month Year>. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ) conducts the Effective Health Care Program as part of its mission to organize knowledge and make it available to inform decisions about health care. As part of the Medicare Prescription Drug, Improvement, and Modernization Act of 2003, Congress directed AHRQ to conduct and support research on the comparative outcomes, clinical effectiveness, and appropriateness of pharmaceuticals, devices, and health care services to meet the needs of Medicare, Medicaid, and the Children's Health Insurance Program (CHIP).

AHRQ has an established network of Evidence-based Practice Centers (EPCs) that produce Evidence Reports/Technology Assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care. The EPCs now lend their expertise to the Effective Health Care Program by conducting comparative effectiveness reviews (CERs) of medications, devices, and other relevant interventions, including strategies for how these items and services can best be organized, managed, and delivered.

Systematic reviews are the building blocks underlying evidence-based practice; they focus attention on the strength and limits of evidence from research studies about the effectiveness and safety of a clinical intervention. In the context of developing recommendations for practice, systematic reviews are useful because they define the strengths and limits of the evidence, clarifying whether assertions about the value of the intervention are based on strong evidence from clinical studies. For more information about systematic reviews, see

<http://www.effectivehealthcare.ahrq.gov/reference/purpose.cfm>

AHRQ expects that CERs will be helpful to health plans, providers, purchasers, government programs, and the health care system as a whole. In addition, AHRQ is committed to presenting information in different formats so that consumers who make decisions about their own and their family's health can benefit from the evidence.

Transparency and stakeholder input from are essential to the Effective Health Care Program. Please visit the Web site (<http://www.effectivehealthcare.ahrq.gov>) to see draft research questions and reports or to join an e-mail list to learn about new program products and opportunities for input. Comparative Effectiveness Reviews will be updated regularly.

We welcome comments on this CER. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director and Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

Steering Committee

Peer Reviewers

<Name>
<Place>
<City>, <ST>

<Name>
<Place>
<City>, <ST>

Structured Abstract

Background: Numerous tools exist to assess methodological quality, or risk of bias in systematic reviews; however, few have undergone extensive reliability or validity testing.

Objectives: 1) assess the reliability of the Cochrane Risk of Bias (ROB) tool for randomized controlled trials (RCTs) and the Newcastle-Ottawa Scale (NOS) for cohort studies between individual raters, and between consensus agreements of individual raters for the ROB tool; 2) assess the validity of the Cochrane ROB tool and NOS by examining the association between study quality and treatment effect size (ES); 3) examine the impact of study-level factors on reliability and validity.

Methods: Two reviewers independently assessed risk of bias for 154 RCTs. For 30 RCTs, two reviewers from each of four Evidence-based Practice Centers assessed risk of bias and reached consensus. Inter-rater agreement was assessed using kappa statistics. We assessed the association between ES and risk of bias: ES were pooled using a random effects model and compared across risk of bias categories. We examined the impact of study-level factors on the association between risk of bias and ES using subgroup analyses. Two reviewers independently applied the NOS to 131 cohort studies from 8 meta-analyses. Inter-rater agreement was calculated using kappa statistics. Within each meta-analysis, we generated a ratio of pooled estimates for each quality domain. The ratios were combined to give an overall estimate of differences in effect estimates with inverse-variance weighting and a random effects model.

Results: Inter-rater reliability between two reviewers was considered fair for most domains (κ ranging from 0.24 to 0.37), except for sequence generation ($\kappa=0.79$, substantial). Inter-rater reliability of consensus assessments across 4 reviewer pairs was moderate for sequence generation ($\kappa=0.60$), fair for allocation concealment and “other sources of bias” ($\kappa=0.37, 0.27$), and poor for the remaining domains (κ ranging from 0.05 to 0.09). Inter-rater variability was influenced by study-level factors including nature of outcome, nature of intervention, study design, trial hypothesis, and funding source. No statistically significant differences were found in ES when comparing studies categorized as high, unclear or low risk of bias; however, trends showed larger ES for studies at high and unclear versus low risk of bias for individual domains. Inter-rater reliability of the NOS varied from substantial for length of followup to poor for selection of non-exposed cohort and demonstration that the outcome was not present at outset of study. We found no association between individual NOS items or overall NOS score and effect estimates.

Conclusion: More specific guidance is needed to apply risk of bias/quality tools. Study-level factors that were shown to influence agreement provide direction for detailed guidance. Low agreement across pairs of reviewers raises questions about the credibility of risk of bias assessments in any given systematic review. This has implications for incorporation of risk of bias into results and grading the strength of evidence. Variable agreement for the NOS, and lack of evidence that it discriminates studies that may provide biased results, challenge its suitability for use in systematic reviews.

Contents

Executive Summary.....	ES-1
Introduction.....	1
Quality and Risk of Bias Assessment in Systematic Reviews.....	1
The Cochrane Risk of Bias Tool	2
The Newcastle-Ottawa Scale.....	2
Goal and Objective	3
Methods.....	5
Steering Committee	5
General Approach	5
Risk of Bias and Randomized Controlled Trials	5
Study Selection.....	5
Risk of Bias Assessments.....	5
Data Extraction.....	6
Data Analysis.....	6
Reliability of the ROB tool.....	6
Validity of the ROB tool.....	7
Newcastle-Ottawa Scale and Cohort Studies.....	7
Study Selection.....	7
Quality Assessments.....	7
Data Extraction.....	8
Data Analysis.....	8
Reliability of the NOS.	8
Validity of the NOS.	8
Results.....	9
Overview	9
Risk of Bias and Randomized Controlled Trials	9
Description of Reviewers	9
Description of Randomized Controlled Trials	9
Inter-rater Reliability	11
Inter-consensus Reliability	12
Validity	12
Newcastle-Ottawa Scale and Cohort Studies.....	14
Description of Reviewers	14
Description of Sample	14
Inter-rater Reliability	14
Validity	15
Summary and Discussion.....	15
Key Points	15
Risk of Bias Tool and Randomized Controlled Trials	15
Newcastle-Ottawa Scale and Cohort Studies	16

Discussion	16
Risk of Bias Tool and Randomized Controlled Trials	16
Newcastle-Ottawa Scale and Cohort Studies	18
Implications for Practice	19
Future Directions	20
Strengths and Limitations	20
Conclusions	21
References	23
Abbreviations	26

Tables

Table 1. Overview of study componentes.....	6
Table 2. Interpretation of Fleiss' kappa (κ)	6
Table 3. Risk of bias assessments by domain (N=154)	10
Table 4. Inter-rater reliability on Risk of Bias assessments, by domain.....	11
Table 5. Inter-rater reliability on Risk of Bias assessments, by domain and study-level variable	11
Table 6. Inter-rater reliability on Risk of Bias assessments across 4 EPCs.....	12
Table 7. Description of meta-analyses of cohort studies included in sample	14
Table 8. Inter-rater reliability on NOS assessments, by domain	14
Table 9. Results of meta-meta-analysis of quality items and measures of association	15
Table 10. Inter-rater reliability on Risk of Bias assessments, comparison across studies.....	17
Table 11. Trials at high or unclear risk of bias across samples	18

Figures

Figure 1. Sequence generation	12
Figure 2. Allocation concealment	12
Figure 3. Blinding	13
Figure 4. Incomplete outcome data.....	13
Figure 5. Selective outcome reporting	13
Figure 6. Other sources of bias	13
Figure 7. Overall risk of bias	13

Appendixes

Appendix A. Steering Committee Members
Appendix B. Sample of Randomized Controlled Trials
Appendix C. Guidelines for Risk of Bias assessments
Appendix D. Variables for data extraction from randomized controlled trials
Appendix E. List of meta-analyses and cohort studies used for NOS assessments
Appendix F. Decision rules for application of the Newcastle-Ottawa Scale
Appendix G. Supplementary Information for NOS Assessments
Appendix H. Description of randomized controlled trials

Executive Summary

Introduction

The assessment of methodological quality, or risk of bias, of studies included in a systematic review (SR) is a key step and serves to: 1) identify the strengths and limitations of the included studies; 2) investigate, and potentially explain, heterogeneity in findings across different studies included in a SR; and, 3) grade the strength of evidence for a given question. There are numerous tools to assess methodological quality, or risk of bias, of primary studies; however, few have undergone extensive inter-rater reliability or validity testing. Moreover, the focus of much of the tool development or testing that has been done has been on criterion or face validity. Therefore it is unknown whether, or to what extent, the summary assessments based on these tools differentiate between studies with likely biased and unbiased results.

There is a need for inter-rater reliability testing of different tools in order to assess and enhance consistency in their application and interpretation across different SRs. Further, validity testing is essential to ensure that the tools being used can identify studies with biased results. Finally, there is a need to determine inter-rater reliability and validity in order to support the uptake and use of individual tools that are being recommended for use by the SR community, and specifically the Cochrane Risk of Bias (ROB) tool within the EPC Program.

Key Questions

The objective of this project was to assess the reliability and validity of quality assessment tools across individual raters and pairs of raters in evaluating study quality in comparative effectiveness reviews and other evidence reports produced through the AHRQ Effective Health Care (EHC) Program. In this work we focused on the Cochrane ROB tool and the Newcastle-Ottawa Scale (NOS). Both are recommended and frequently used in systematic reviews of randomized controlled trials (RCTs) and cohort studies, respectively.

The specific objectives were:

1. To assess the reliability of the Cochrane ROB tool for RCTs and the NOS for cohort studies between individual raters and, for the ROB tool, between the consensus agreements of individual raters (i.e., comparing consensus agreements across four EPCs).
2. To assess the validity of the Cochrane ROB tool and NOS by using empirically-shown evidence of inverse association between risk of bias or study quality and effect size (ES) as a construct for validity.
3. To examine the impact of study-level factors (e.g., outcomes, interventions and conditions) on scale reliability and validity.

Methods

Cochrane Risk of Bias Tool and Randomized Controlled Trials

Study Selection: A sample of 154 RCTs involving adults was randomly selected from among 616 trials published in December 2006 that were examined for quality of reporting.¹ As the parameters required for sample size calculations in this type of work are presently unknown, we used a pragmatic approach to determine sample size. This was based on previous studies in this area, input from the Steering Committee, and the availability of resources and timelines. Hence, we selected a 25 percent random sample of the 616 trials described above.

Risk of Bias Assessments: We pilot tested the ROB tool and developed decision rules to accompany the guidance for applying the tool that is publicly available in the Cochrane Handbook.² The tool was applied to each study independently by two reviewers. For each study, one reviewer was from the xx EPC and one from the xx EPC. To assess reliability between consensus agreements, we used a subset of 30 trials. Two reviewers at each of the four collaborating EPCs independently assessed risk of bias and reached consensus (xx, xx, xx, xx).

Data Extraction: We extracted data on the primary outcome for each trial. Several characteristics of the trial that may also be related to risk of bias were extracted, including study type (efficacy, equivalence), study design (parallel, crossover), the condition being treated, type of outcome (subjective, objective), nature of the intervention (pharmacological, nonpharmacological), treatment mode (flexible dose vs. fixed dose), treatment duration, baseline mean difference between study groups for continuous outcomes, the impact of the intervention (treatment ES), variance in ES, sample size, and funding source. Data extraction for each study was completed at the xx EPC by a single reviewer. A 10 percent random sample of trials was checked by a second reviewer.

Data Analysis: For the entire sample of trials, inter-rater agreement between two reviewers was calculated for each domain using weighted kappa statistics. Agreement was categorized as poor, slight, fair, moderate, substantial, or almost perfect using accepted approaches (Table ES-1).³ Using subgroup analyses, we explored whether inter-rater agreement was influenced by study-level factors, including study design, study hypothesis, nature of the intervention, nature of the outcome, and source of funding. For the subset of 30 studies, agreement for consensus assessments across pairs of reviewers was measured using unweighted kappa statistics (i.e., the consensus assessments were compared across the pairs of reviewers from 4 EPCs).

Since there is no gold standard against which the validity of the ROB assessments can be made, the empirically shown inverse association between the ES and the study quality based on ROB assessments was taken as construct validity. For each RCT we calculated an ES for the primary outcome. ES were calculated using Cohen's d for continuous outcomes. Odds ratios were calculated for dichotomous outcomes and converted into ES using a method developed by Hasselblad and Hedges.⁴ The ES for all RCTs were combined using a random effects model.⁵ We compared the pooled ES for the high, unclear, and low risk of bias categories for each of the six domains and overall risk of bias. The differences were compared statistically using Kruskal-Wallis and Spearman tests.

The effect of specific covariates on risk of bias was analyzed using logistic regression. We also tested these covariates for their effect on the association between risk of bias and ES in a subgroup analysis. The covariates examined were intervention type (pharmacological or nonpharmacological), study design (parallel vs. other), funding source (pharmaceutical industry

vs. other), type of trial (efficacy/superiority vs. other), and type of outcome (subjective or objective).

Table ES-1. Interpretation of Fleiss' kappa (κ)(from Landis and Koch 1977)³

K	Interpretation
<0	Poor agreement
0.0-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.0	Almost perfect agreement

Newcastle-Ottawa Scale and Cohort Studies

Study selection: We used an iterative approach to identify a sample of cohort studies based on already completed meta-analyses of cohort studies. Initially, we searched completed EPC reports. We found 3 EPC reports with relevant meta-analyses including 36 cohort studies that met the inclusion criteria (see below). We subsequently conducted searches in Medline using the terms to capture systematic reviews (meta-analys?s.mp, review.pt and search.tw), cohort studies (exp Cohort Studies/, cohort\$.tw, (observation\$ adj stud\$).tw) and meta-analyses (exp meta-analysis/, (analysis adj3 (group\$ or pool\$)).tw, (forest adj plot\$).mp). Results were limited to studies in humans, English language reports, and those published in 2000 or later. We searched by year starting with the most recent, and continued until we identified a sufficient number of studies.

A meta-analysis was considered appropriate to include if it incorporated at least 10 studies, assessed a dichotomous outcome, and had substantial statistical heterogeneity (i.e., $I^2 > 50$ percent). Previous meta-epidemiological research has used a minimum sample size per meta-analysis of 5 to 10 studies.^{6,7} This ensures that there is a sufficient pool of studies with some degree of variability in each meta-analysis in order to test the hypotheses. Some degree of heterogeneity is required in order to test whether quality as assessed by the NOS tool can differentiate studies with different effect estimates.

Our target sample size was 125 cohort studies from appropriate meta-analyses. Our final sample included 131 cohort studies from 8 meta-analyses.

Quality Assessments: We pilot tested the NOS tool and developed decision rules to accompany existing guidance for the NOS. All studies were assessed using the NOS independently by two reviewers. One reviewer was from the xx EPC and one reviewer was from the xx EPC. Discrepancies were resolved through discussion to produce consensus assessments for each study.

Data Extraction: The outcomes and data for effect estimates were based on the meta-analysis and checked against the primary studies by a single reviewer. The statistician double-checked data that were unclear.

Data Analysis: Inter-rater agreement was calculated for each domain and for overall quality assessment using weighted or unweighted kappa statistics, as appropriate. Agreement was categorized as above (Table ES-1). For the results of the individual meta-analyses, we coded endpoints consistently so that the outcome occurrence was undesired. Within each meta-analysis, we generated a ratio of odds ratios (i.e., odds ratios for studies with and without the domain of

interest or of high/low quality as assessed by the NOS). To maintain consistency, we used odds ratios to summarize all meta-analyses, even if this was not the statistic that was used in the original meta-analysis. The ratios of odds ratios for each meta-analysis were combined to give an overall estimate of differences in effect estimates using meta-analytic techniques with inverse-variance weighting and a random effects model.⁸

Results

Results are presented according to the tools we examined: ROB tool for RCTs and NOS for cohort studies.

Risk of Bias and Randomized Controlled Trials

Description of reviewers: Twelve reviewers from the xx EPC and the xx EPC assessed the RCTs using the ROB tool. Individuals had varying levels of relevant training, experience with systematic reviews in general, and experience with EPC work specifically. The length of time they had worked with their respective EPC ranged from 9 months to 10 years. Ten of the 12 reviewers had formal training in systematic reviews. Three of the reviewers had a doctoral degree in epidemiology or health/clinical sciences; 8 reviewers had a master's degree in epidemiology/public health, health/clinical sciences, or math/statistics; and one reviewer had an undergraduate degree in health sciences.

For the subset of 30 RCTs, two reviewers from each of the four EPCs were involved in applying the ROB tool and reaching consensus for each study. The reviewers had the following backgrounds: PhD (n=4), MD (n=1), PhD students with completed master's degrees (n=2), MD and Master's degree (n=1), and Master's degree (n=1). The length of time they had worked with an EPC ranged from 2 to 10 years. Six reviewers had formal training in SRs.

Description of sample: We included 154 RCTs. The majority of trials were published in specialty medical journals (87.7 percent). The median impact factor of the journal was 2.9 (interquartile range [IQR] 1.8, 5.1) and the mean number of authors was 6.8 (standard deviation 3.3). The country represented most frequently was the United States (31.8 percent). Approximately 70 percent of trials declared a funding source with industry (27.3 percent) and government (26.0 percent) sources most frequent.

The design of the majority of trials was parallel (81.8 percent), efficacy/superiority (84.4 percent) with individuals as the unit of randomization (95.5 percent). Just over half of the trials examined drug interventions (53.3 percent), with behavioral/psychological (11.0 percent) and surgical (11.7 percent) interventions commonly represented. The median sample size was 63 (IQR 39, 123).

A wide range of diagnostic categories was represented. These were classified into Aging; Cancer Research; Circulatory and Respiratory Health; Gender and Health; Genetics; Health Services and Policy Research; Human Development, Child and Youth Health; Infection and Immunity; Musculoskeletal Health and Arthritis; Neurosciences, Mental Health and Addiction; Nutrition, Metabolism and Diabetes; and Population and Public Health. The most frequently represented categories were Circulatory and Respiratory Health (18.2 percent), Nutrition, Metabolism and Diabetes (17.5 percent), and Musculoskeletal Health and Arthritis (14.9 percent). The primary outcomes were objective in 48.1 percent of trials and subjective in 51.9 percent. Source of outcome assessment was primarily by clinician (35.1 percent), laboratory measure (23.4 percent), or self-report (23.4 percent).

The vast majority of trials had overall risk of bias assessments of high (46.8 percent) or unclear (52.6 percent) with only one trial assessed as low risk of bias overall (0.7 percent). Table ES-2 provides details on the risk of bias assessments for the individual domains. The domains that were most frequently rated as low risk of bias were sequence generation (54.6 percent), missing outcome data (63.6 percent), and selective reporting (77.3 percent). The remaining domains were most frequently assessed as unclear: allocation concealment (77.3 percent), blinding (48.7 percent), and “other sources of bias” (55.8 percent). These results should be interpreted with caution given the low level of agreement between reviewers (Table ES-3).

We explored study-level variables and their association with domain-specific and overall risk of bias. *Sequence generation* was influenced by the nature of the outcome (objective outcomes showed higher risk of bias, $p=0.01$) and study design (parallel showed lower risk of bias, $p=0.02$). *Allocation concealment* was also influenced by the nature of the outcome with objective outcomes having higher risk of bias ($p=0.0007$). *Blinding* was influenced by nature of the intervention with pharmaceutical interventions having lower risk of bias ($p=0.01$). *Selective outcome reporting* was associated with the nature of the intervention (surgical trials showed higher risk of bias, $p=0.002$) and funding (industry support had higher risk of bias, $p=0.04$). “*Other sources of bias*” was associated with funding with industry funding showing higher risk of bias ($p<0.0001$). Overall risk of bias was also associated with funding with industry funding showing higher risk of bias ($p<0.0001$). Of note, “other sources of bias” incorporates several considerations including “inappropriate influence of the study sponsor” (i.e., the extent and nature of involvement of the study sponsor and whether this would likely lead to biased results) which is different from source of funding (i.e., whether the study was funded by industry).

Table ES-2. Risk of bias assessments by domain* (N=154)

Domain	Risk of bias assessments – n (%)		
	High	Unclear	Low
Sequence generation	0 (0.0)	70 (45.5)	84 (54.6)
Allocation concealment	2 (1.3)	119 (77.3)	33 (21.4)
Blinding	21 (13.6)	75 (48.7)	58 (37.7)
Incomplete data	29 (18.8)	27 (17.5)	98 (63.6)
Selective reporting	16 (10.4)	19 (12.3)	119 (77.3)
Other sources of bias	33 (21.4)	86 (55.8)	35 (22.7)
Overall risk of bias	72 (46.8)	81 (52.6)	1 (0.7)

* The risk of bias assessments presented here are based on consensus between two reviewers.

Inter-rater reliability: Inter-rater reliability for the RCTs is presented by domain in Table ES-3. Sequence generation had the highest level of agreement, which was considered substantial. Reliability for the remaining domains was fair.

A random sample of 30 studies was selected to compare consensus assessments across pairs of reviewers from the four participating EPCs. The results of the inter-EPC reliability are detailed in Table ES-3. There was moderate agreement for sequence generation, fair agreement for allocation concealment and “other sources of bias,” and slight agreement for the remaining domains and overall risk of bias.

Table ES-3. Inter-rater reliability on Risk of Bias assessments, by domain

Domain	Between 2 reviewers (n=154)		Between pairs of reviewers (n=30)	
	Agreement (weighted κ)	Interpretation (³)	Agreement (κ)	Interpretation (³)
Sequence generation	0.79	Substantial	0.60	Moderate

Allocation concealment	0.24	Fair	0.37	Fair
Blinding	0.33	Fair	0.09	Slight
Incomplete data	0.34	Fair	0.05	Slight
Selective reporting	0.27	Fair	0.08	Slight
Other sources of bias	0.24	Fair	0.27	Fair
Overall risk of bias	0.26	Fair	0.10	Slight

We assessed whether important study-level variables influenced inter-rater reliability. Assessments for *sequence generation* and *incomplete outcome data* were not influenced by any variable. For *allocation concealment*, inter-rater agreement was better for trials with parallel (0.32, fair) versus other designs (-0.07, poor) ($p=0.0002$) and for those without (0.38, fair) versus with (-0.10, poor) industry funding ($p=0.03$). In terms of *blinding*, inter-rater agreement was better for objective (0.54, moderate) versus subjective (0.18, slight) outcomes ($p=0.02$), and trials with other (0.77, substantial) versus parallel (0.27, fair) designs ($p=0.0004$). For *selective outcome reporting*, inter-rater agreement was greater for trials with hypotheses of efficacy/superiority (0.38, fair) versus others (e.g., equivalence, non-inferiority; -0.31, poor) ($p<0.0001$). For “*other sources of bias*,” inter-rater agreement was better for trials examining pharmacological (0.38, fair) versus nonpharmacological (-0.06, poor) interventions ($p=0.02$), and subjective (0.45, moderate) versus objective (0.09, slight) outcomes ($p=0.04$).

Validity: No statistically significant differences were found in ES across the domain-specific and overall risk of bias categories. In five of the seven cases, studies in the high risk of bias category had average ES greater than studies that were low risk of bias. The exceptions were allocation concealment and “other sources of bias.” In six of the seven cases, studies at unclear risk of bias had average ES greater than studies at low risk of bias; in the remaining case (incomplete outcome data) the ES were the same. There was no impact when controlling for study-level factors (i.e., no statistically significant differences were found).

Newcastle-Ottawa Scale and Cohort Studies

Description of Reviewers: Sixteen reviewers from the xx EPC and the xx EPC assessed the studies using the NOS. Individuals had varying levels of relevant training, experience with systematic reviews in general, and experience with EPC work specifically. The length of time they had worked with their respective EPC ranged from 4 months to 10 years. Thirteen reviewers had formal training in systematic reviews. Four reviewers had a doctoral degree; 10 reviewers had a master’s degree; 1 reviewer had a medical degree and master’s degree; and 1 reviewer had an undergraduate degree.

Description of Sample: The 131 cohort studies were taken from 8 meta-analyses which covered a variety of topics: breastfeeding and asthma ($n=10$ studies); impaired glucose tolerance and diabetes mellitus ($n=17$); cardiac resynchronization therapy and all-cause mortality ($n=11$); drug-resistant tuberculosis and positive treatment outcome ($n=17$); statins and mortality from severe infections and sepsis ($n=20$); red meat intake and prostate cancer ($n=15$); overweight and obesity and preterm birth before 37 weeks ($n=38$); and antenatal depression and preterm birth ($n=20$).

Inter-rater reliability: Inter-rater reliability for the 131 cohort studies is presented by domain in Table ES-4. The item “was the followup long enough for the outcome to occur” had the highest level of agreement which was considered substantial. Reliability was moderate for *ascertainment*

of exposure and ascertainment of outcome. Reliability was fair for representativeness of the cohort, and slight for comparability of cohorts and adequacy of followup of cohorts. Selection of the non-exposed cohort and demonstration that the outcome was not present at the outset of the study had poor reliability. Reliability for the overall score (total number of stars) was fair.

Table ES-4. Inter-rater reliability on NOS assessments, by domain

Domain	Agreement (κ)*	Interpretation ³
Representativeness of the exposed cohort	0.23	Fair
Selection of the non-exposed cohort	-0.03	Poor
Ascertainment of exposure	0.43	Moderate
Demonstration that the outcome was not present at outset of study	-0.06	Poor
Comparability	0.18	Slight
Assessment of outcome	0.49	Moderate
Length of follow-up sufficient	0.68	Substantial
Adequacy of participant followup	0.29	Fair
Total stars	0.29*	Fair

NA=not applicable

* We used a weighted kappa for the total score as it assumes some ordinality in the assessment; other kappas are not weighted, i.e., Cohen's kappa.

Validity: We found no association between individual NOS items or overall NOS score and effect estimates.

Summary and Discussion

Summary Points

Risk of Bias Tool and Randomized Controlled Trials:

- Inter-rater reliability between reviewers was fair for all domains except sequence generation, which was substantial.
- Inter-rater reliability between pairs of reviewers was moderate for sequence generation, fair for allocation concealment and “other sources of bias,” and slight for the remaining domains.
- Low agreement between reviewers suggests the need for more specific guidance regarding interpretation and application of the ROB tool or possibly re-phrasing of items for clarity.
- Examination of study-level variables and their association with inter-rater agreement identified areas that require specific guidance in applying the ROB tool. For example, nature of the outcome (objective vs. subjective), study design (parallel vs. other), and trial hypothesis (efficacy/superiority vs. other).
- Low agreement between pairs of reviewers indicates the potential for inconsistent application and interpretation of the ROB tool across different groups and systematic reviews.
- Most RCTs in the sample were assessed as high or unclear risk of bias for many domains. This raises concerns about the methodological rigor of trials in general, and the ability of the ROB tool to detect differences across trials that may be associated with biases in estimates of treatment effects.

- No statistically significant differences were found in ES across high, unclear, and low risk of bias categories; however, trends consistently showed greater effect estimates for studies at high or unclear risk of bias.

Newcastle-Ottawa Scale and Cohort Studies:

- Inter-rater reliability between reviewers ranged from poor to substantial, but was poor or fair for the majority of domains.
- No association was found between individual quality domains and measures of association.

Discussion

Risk of Bias Tool and Randomized Controlled Trials: We found that inter-rater reliability between reviewers was low for all but one domain in the ROB tool. These findings are similar to results of previous research.⁹ The sample of trials used in this study was not part of a systematic review, rather they were trials randomly selected from a larger pool. Hence, the trials covered a wide range of topics. This may have contributed to some of the low agreement as reviewers had to consider different nuances for each trial. Previous research has demonstrated greater agreement within the context of a systematic review where all trials examined the same interventions in similar populations.¹⁰ Nevertheless, the low agreement raises concerns, and points to the need for clear and detailed guidance in terms of applying the ROB tool. One of the unique contributions of the present study was the analysis of inter-rater reliability controlling for study-level variables. This provides some direction for where more specific guidance may be beneficial. For instance, agreement was considerably lower for: allocation concealment when trials did not have a parallel design; blinding when the nature of the outcome was subjective; selective outcome reporting when the trial hypothesis was not one of efficacy/superiority; and “other sources of bias” for nonpharmacological interventions and when the outcome was subjective. In summary, agreement may be better in classic parallel trials of pharmacological interventions, whereas trials with different design features (e.g., crossover) or hypotheses (e.g., equivalence, non-inferiority), and those examining nonpharmacological interventions appear to introduce more ambiguity for risk of bias assessments.

Another unique contribution of the present study was the examination of the consensus ratings across pairs of reviewers. These ratings should be free of individual rater errors and bias given that these are combined ratings with disagreements resolved. Further, this is a more meaningful measure of agreement (as opposed to reliability between two reviewers), as these ratings are the ones reported in systematic reviews. In this study, the pairs of reviewers were from four different centers, each with a long history of producing systematic reviews. The agreement across the pairs of reviewers was generally lower than the agreement between reviewers. This raises concerns about the variability in interpreting and applying the ROB tool that can occur across different groups and across systematic reviews. It also raises questions regarding the credibility of the risk of bias assessments within any given systematic review.

Overall risk of bias was high or unclear in 99 percent of the studies used for this research. This is consistent with other studies where the vast majority of trials have studies assessed as high or unclear risk of bias overall. This raises the question of whether all these trials are in fact substantially flawed or whether the ROB tool is overly punitive. If the vast majority of trials are assessed as high or unclear risk of bias, the ROB tool may not be sensitive to differences in

methodology that might explain variation in treatment effect estimates across studies (e.g., study methodology as a potential explanation for heterogeneity in meta-analyses). Questions also arise regarding whether poor assessments are a result of inadequate or unclear reporting at the trial level. While the focus of the ROB tool is intended to be on methods rather than reporting, reviewers regularly indicate that they rely on the trial reporting to make their assessments. Even within recent samples of trials that were published after the emergence and widespread dissemination of reporting guidelines, we see high proportions assessed as high or unclear risk of bias. The risk of bias assessments were less severe within the individual domains. However, for the current sample most trials were assessed as high or unclear risk of bias for three of the six domains, including allocation concealment, blinding, and “other sources of bias.” These findings may be beneficial for developers and promoters of reporting guidelines, as well as for researchers who are reporting randomized trials.

We found no statistically significant association between effect estimates and risk of bias assessments. The main explanations for this finding are that either there is no association, or more likely, there was insufficient power to detect differences. One of the factors contributing to low power was the small number of studies within certain domains in the low risk of bias category. This was particularly the case for overall risk of bias as there was only one study in the low category. However, the trend was evident in that the studies at high and unclear risk of bias overall had substantially greater treatment ES (ES=0.94 and 0.85, respectively vs. 0.31). The trend for five of the seven domains (including overall risk of bias) was for greater treatment effect estimates for studies at high risk of bias compared to low risk of bias. Further, in all but one domain, studies at unclear risk of bias had greater treatment effect estimates than studies at low risk of bias, although this was not statistically significant. This finding is important in interpreting evidence: when risk of bias is unclear, estimates are likely to be overestimating treatment effects.

Newcastle-Ottawa Scale and Cohort Studies: This is the first study to our knowledge that has examined inter-rater reliability and construct validity of the NOS. We found a wide range in the degree of agreement across the domains of the NOS, ranging from slight to substantial. The domain with substantial agreement was not surprising. This domain asked “was the followup long enough for the outcome to occur?” A priori we asked clinical experts to provide the minimum length of followup for each review question. Thus, the assessors had very specific guidance for this item. The agreement for ascertainment of exposure and assessment of outcome was moderate, suggesting that the wording and response options are reasonable. The remaining items had poor, slight, or fair agreement.

In general, the reviewers found the tool difficult to use and found the decision rules vague even with the additional information we provided as part of this study. General points that arose were whether to assess each study based on the individual report, or as it related to the systematic review question – for example, a cohort study might have reported/assessed comparability between exposed and nonexposed that it was designed to investigate, but also reported (subgroup) outcome data, without reporting corresponding baseline comparability data, by presence or absence of a covariate determining exposure and nonexposure for the meta-analysis of interest. Similarly, there was uncertainty about whether to base assessments on the information contained in the specific study report, or whether to incorporate information from other reports of the same study.

Response options on the NOS caused discordance among reviewers. They found it difficult to determine the difference between some response options (e.g., “truly” vs. “somewhat” representative study population). Furthermore, the importance of the distinction between certain categories was unclear. In some domains multiple responses garnered a star (i.e., a point in the overall score), hence there was no difference in the final score. Reviewers experienced difficulty in interpreting the terminology (e.g., “selected” population) and in some cases the differences between categories were difficult to distinguish (e.g., “structured interview” vs. “written self-report”).

Reviewers also expressed uncertainty regarding the item assessing comparability, unsure whether to indicate that the study controlled for a given confounder if it was not included in the final model due to lack of significance in preliminary analyses. Reviewers expressed uncertainty regarding what some of the domains actually measured (e.g., selection bias vs. applicability). Further, some concerns were raised that the response categories within a domain measured different constructs.

Reviewers commented that they would have liked “unclear” or “no description” options for some of the items.

We found no association between NOS items and the measures of association using meta-epidemiological methods that control for heterogeneity due to condition and intervention. Moreover, we saw no trends suggesting an association between magnitude of association and quality.

Implications for Practice

The findings of this research have critical implications for practice and the interpretation of evidence. The low level of agreement between reviewers and pairs of reviewers puts into question the credibility or validity of risk of bias/quality assessments made with the ROB tool and the NOS within any given systematic review. Moreover, in measurement theory, reliability is a necessary condition for validity (i.e., without being reliable a test cannot be valid). Systematic reviewers are urged to incorporate considerations of risk of bias/quality into their results. Furthermore, integration of the GRADE tool into systematic reviews necessitates the consideration of risk of bias/quality assessments in rating the strength of evidence and ultimately recommendations for practice. The results and their interpretation in a systematic review will be misleading if they are based on flawed assessments of risk of bias/quality. Moreover, variability across reviewers and review groups may produce arbitrary results.

There is an urgent need for more detailed guidance to apply these tools. In the meantime, reviewers and review teams need to be aware of the limitations of existing tools. Detailed guidelines, decision rules, and transparency are needed so that readers and end-users of systematic reviews can see how the tools were applied. Further, pilot testing and development of review-specific guidelines and decision rules should be mandatory and reported in detail.

This study provides some evidence of association (or trends) between risk of bias domains of the ROB tool and estimates of treatment effect, which corroborates previous findings. The results confirm that the tool is doing what it is intended to do, i.e., identifying studies that may yield less reliable estimates of treatment effects. We did not find similar evidence for the NOS. Further, the NOS in its current form does not appear to provide reliable quality assessments and requires further development and more detailed guidance. The NOS was previously endorsed by The Cochrane Collaboration; however, more recently the Collaboration has proposed a modified ROB tool to be used for nonrandomized studies. A new tool developed through the EPC Program

for quality assessment of nonrandomized studies offers another alternative. These tools warrant further evaluation.

Future Research

There is a dire need for more detailed guidelines to apply both the ROB and NOS tools, as well as revisions to the tools to enhance clarity. We have identified specific trial features for which clearer guidance is needed. A living database that collects examples of risk of bias/assessments and consensus from a group of experts would be a valuable contribution to this field. Individual review teams and research groups should be encouraged to begin identifying examples and these could be compiled across programs (e.g., the EPC Program) and entities (e.g., The Cochrane Bias Methods Group), and made widely accessible. We have identified specific problems with application and interpretation of the NOS tool. Further revisions and guidance are needed to support the continued use of NOS in systematic reviews. Investment in further reliability and validity testing of other tools is warranted (e.g., Cochrane ROB tool for nonrandomized studies, EPC quality assessment tool). Finally, consensus in this field is needed in terms of the threshold for inter-rater reliability of a measurement before it can be used for any purpose, even descriptive purposes (i.e., describing the risk of bias or quality of a set of studies).

Strengths and Limitations

This is one of few studies examining the reliability and validity of the ROB tool. It is the first to our knowledge that examines reliability between the consensus assessments of pairs of reviewers. Further, it is the first study to provide empirical evidence on study-level variables that may impact reliability of ROB assessments. This is the first study to our knowledge to examine the reliability and validity of the NOS.

The main limitation of the research is that the sample sizes (154 RCTs, 131 cohort studies) may not have provided sufficient power to detect statistically significant differences in effect estimates according to risk of bias/quality. We observed trends for RCTs, with larger effect estimates for studies at high or unclear versus low risk of bias. We found no significant associations between quality and measures of association within the cohort studies, which could be attributable to low power. Furthermore, we did not find any discernable trends. We specifically selected meta-analyses with substantial heterogeneity in order to optimize our potential to see whether quality as assessed with the NOS might explain variations in measures of association.

We involved a number of reviewers with different levels of training, type of training, and extent of experience in quality assessment and systematic reviews. Some of the variability or low agreement may be attributable to characteristics of the reviewers. Nevertheless, all reviewers had previous experience in systematic reviews and quality assessments, and likely represent the range of individuals that would typically be involved in these activities within a systematic review.

A final caveat to note is that the ROB tool has undergone some revisions since we initiated the study. These are detailed in the most recent version of the Cochrane Handbook but were not incorporated into our research. The changes affected primarily the blinding and the “other sources of bias” domains. This does not impact the general findings from our research; however, further testing with the modified tool is warranted.

Conclusions

More specific guidance is needed to apply and interpret risk of bias/quality tools. We identified a number of study-level factors that influence agreement. This information provides direction for more detailed guidance. Low agreement across pairs of reviewers raises questions about the credibility of risk of bias assessments in any given systematic review. This has implications for incorporation of risk of bias into results and grading the strength of evidence. There was variable agreement across items in the NOS. This finding, combined with a lack of evidence that it discriminates studies that may provide biased results, challenges its suitability for use in systematic reviews.

Introduction

Quality and Risk of Bias Assessment in Systematic Reviews

The internal validity of a study reflects the extent to which the design and conduct of the study have prevented bias(es).¹¹ One of the key steps in a systematic review is assessment of a study's internal validity, or potential for bias. This assessment serves to: 1) identify the strengths and limitations of the included studies; 2) investigate, and potentially explain heterogeneity in findings across different studies included in a systematic review; and 3) grade the strength of evidence for a given question.

With the increase in the number of published systematic reviews¹² and development of systematic review methodology over the past 15 years,¹¹ close attention has been paid to the methods for assessing internal validity. Until recently this has been referred to as “quality assessment” or “assessment of methodological quality.”¹¹ In this context “quality” refers to “the confidence that the trial design, conduct, and analysis has minimized or avoided biases in its treatment comparisons.”¹³ To facilitate the assessment of methodological quality, a plethora of tools has emerged.¹³⁻¹⁶ Some of these tools were developed for specific study designs (e.g., randomized controlled trials (RCTs), cohort studies, case-control studies), while others were intended to be applied to a range of designs. The tools often incorporate characteristics that may be associated with bias; however, many tools also contain elements related to reporting (e.g., was the study population described) and design (e.g., was a sample size calculation performed) that are not related to bias.¹¹ The Cochrane Collaboration recently developed a new tool to assess the potential risk of bias in RCTs. The Risk of Bias (ROB) tool¹¹ was developed to address some of the shortcomings of existing quality assessment instruments, including over-reliance on reporting rather than methods.

While there are numerous tools to assess methodological quality, or risk of bias of primary studies,^{11,16,17} few have undergone extensive inter-rater reliability or validity testing. Moreover, the focus of much of the tool development or testing that has been done has been on criterion or face validity.^{11,16,17} Therefore it is unknown whether, or to what extent, the summary assessments based on these tools differentiate between studies with biased and unbiased results (i.e., studies that may over- or underestimate treatment effects).

There is a clear need for inter-rater reliability testing of different tools in order to enhance consistency in their application and interpretation across different systematic reviews. Further, validity testing is essential to ensure that the tools being used can identify studies with biased results. Finally, there is a need to determine inter-rater reliability and validity in order to support the uptake and use of individual tools that are recommended by the systematic review community, and specifically the ROB tool within the Evidence-based Practice Center (EPC) Program.

In this project we focused on two tools that are commonly used in systematic reviews. The Cochrane ROB tool was designed for RCTs and is the instrument recommended by The Cochrane Collaboration for use in systematic reviews of RCTs. The Newcastle-Ottawa Scale is commonly used for nonrandomized studies, specifically cohort and case-control studies. It has also been endorsed for use in systematic reviews of nonrandomized studies by The Cochrane Collaboration. In the sections that follow we describe these tools, their development, and any testing that has occurred.

The Cochrane Risk of Bias Tool

The Cochrane Collaboration released a new tool in 2008 to assess the potential risk of bias in RCTs.¹¹ The original ROB tool was based on six domains: sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, and “other sources of bias” (e.g., design-specific risks of bias; early stopping for benefit; severe baseline imbalances; inappropriate influence of funders). The developers of the tool aimed to distinguish between actual methods of conducting the trials versus reporting. Furthermore, the choice of components for inclusion in the tool was based on empirical evidence demonstrating their association with effect estimates. There is a growing body of evidence from methodological studies, and meta-epidemiological studies in particular, to quantify the extent to which different characteristics of a trial exaggerate treatment effects. Empirical evidence exists for the following characteristics: sequence generation; allocation concealment; blinding; incomplete outcome reporting; selective outcome reporting; and, inappropriate influence of the funder.² In 2011, The Cochrane Collaboration released a new version of the ROB tool which incorporated modifications based on user testing and feedback.²

Researchers at the University of Alberta Evidence-based Practice Center (UAEPC) evaluated the original Cochrane ROB tool in a sample of trials with a number of treatment conditions and showed that inter-rater agreement ranged from slight to substantial across the different domains, with the overall risk of bias assessment having ‘fair’ agreement.⁹ The authors further showed that treatment effect sizes (ES) differed: studies at high or unclear risk of bias reported significantly greater treatment effects (ES=0.52) than those at low risk of bias (ES=0.23). The authors identified sources of discrepancy and made recommendations in order to enhance the degree of consistency of the ROB tool. One of the stated limitations of this research was that the sample to which the tool was applied included only trials in children, the results of which may not be generalizable to trials conducted in other populations. A subsequent study by the same researchers showed improved inter-rater agreement on ROB assessments within the context of a systematic review.¹⁰ The authors suggested that the improved agreement may have resulted from review-specific guidelines and pilot-testing. No important patterns appeared in analyses comparing effect estimates and risk of bias; however, the ES were very homogeneous across the studies and there were very few studies in the sample that were at low risk of bias. This may have led to inadequate power to detect differences.

The Newcastle-Ottawa Scale

The Newcastle-Ottawa Scale (NOS) is a quality assessment tool for use on nonrandomized studies included in systematic reviews, specifically cohort and case-control studies. The tool was produced by the combined efforts of the Universities of Newcastle, Australia and Ottawa, Canada¹⁸ and was first reported at the 3rd Symposium for Systematic Reviews in Oxford, UK in 2000.¹⁹ Separate assessment criteria are available for case-control and cohort studies, and evaluate: the selection of participants, comparability of study groups, and the ascertainment of exposure (case-control studies) or outcome of interest (cohort studies). A star rating system is used to indicate the quality of a study, with a maximum assessment of nine stars.¹⁹ Each criterion receives a single star if appropriate methods have been reported. The selection domain is subdivided to evaluate the selection of the exposed and non-exposed cohorts, the ascertainment of exposure, and whether the study demonstrated that the outcome of interest was not present at the start of the study. Comparability is the only category that may receive two stars: one if the most important confounders have been adjusted for in the analysis, and a second star if any other

adjustments were made. Outcome of interest is made up of three questions: the appropriateness of the methods used to evaluate the outcome, the length of followup, and the degree of the loss to followup.¹⁸

The developers of the NOS have examined face and criterion validity, inter-rater reliability, and evaluator burden for the NOS. Face validity has been evaluated as strong by comparing each individual assessment item to their stem question. Criterion validity has shown a strong agreement with the Downs and Black assessment tool²⁰ on a series of 10 cohort studies evaluating hormone replacement therapy in breast cancer, with an intra-class correlation (ICC) of 0.88. Inter-rater reliability for the NOS on cohort studies was high with an ICC of 0.94. Evaluator burden, as assessed by the time required to complete the NOS evaluation, was shown to take significantly less time than the Downs and Black tool ($p < 0.001$).²¹ The authors state that further assessment of the construct validity and the relationship between the external criterion of the NOS and its internal structures are under consideration.¹⁸ These studies have been presented as abstracts; at present no peer reviewed articles have been published investigating the psychometric properties of the NOS.

Goal and Objective

We undertook this project to assess the reliability and validity of the two tools described above. We were interested in the reliability of risk of bias/quality assessments across individual raters, and between consensus agreements of individual raters. This work is directly relevant to the methods and interpretation of data in comparative effectiveness reviews and other evidence reports produced through the AHRQ Effective Health Care (EHC) Program. This project was done in collaboration with the xx EPC, xx EPC, and xx EPC.

The specific objectives were:

1. To assess the reliability of the Cochrane ROB tool for RCTs and the NOS for cohort studies between individual raters and, for the ROB tool, between the consensus agreements of individual raters (i.e., comparing consensus agreements across four EPCs).
2. To assess the validity of the Cochrane ROB tool and NOS by using empirical evidence of inverse association between risk of bias or study quality and ES as a construct for validity.
3. To examine the impact of study-level factors (e.g., outcomes, interventions, and conditions) on scale reliability and validity.

Methods

Steering Committee

A steering committee provided direction to the individual components of the project. The committee provided a similar function as the technical expert panel in evidence reports. Members of the Steering Committee are listed in Appendix A.

General Approach

We developed a protocol that detailed our methods prior to the start of the study. The protocol was reviewed by the Steering Committee and approved by AHRQ.

We proposed two different statistical approaches to assess validity against the treatment effect size (ES), which we consider as construct validity. The first approach is based on effect estimates from primary studies, while the second was a meta-epidemiologic approach which controls for confounding and heterogeneity due to study-level factors (e.g., methodology, outcomes, interventions/exposures, and conditions).

Risk of Bias and Randomized Controlled Trials

Study Selection

A sample of 154 recently conducted RCTs involving adults was randomly selected from a convenience sample of 616 trials published in December 2006. These trials were previously examined for quality of reporting by Hopewell and colleagues (Appendix B).¹ We chose this sample as it presented several advantages including efficiencies in sample identification, as well as the potential for validation of assessments for key variables by comparing them with those of another independent study team.

Conducting sample size calculations for a meta-analysis is challenging and cannot be determined using standard approaches to sample size calculations done for other research designs, such as RCTs. There are a number of parameters required for sample size calculations that are presently unknown for research of this nature. Therefore, we used a pragmatic approach to determine sample size. This was based on previous studies in this area, input from the Steering Committee, and the availability of resources and timelines. We chose to select a 25 percent random sample of the 616 trials described above.

Risk of Bias Assessments

The ROB tool was applied to each study independently by two reviewers who had training and experience with the tool. A pool of reviewers was assembled from staff at the xx EPC and xx EPC. To assess reliability between consensus agreements, we used a subset of 30 trials. Two reviewers at each of the four collaborating EPCs independently assessed risk of bias and reached consensus (xx EPC, xx EPC, xx EPC, xx EPC). Table 1 provides an overview of the number of reviewers and number of studies for each component of this study.

All reviewers involved in the project pilot tested the ROB tool. We applied the tool to five trials and met by teleconference to discuss any disagreements in general interpretation of the tool. Decision rules were developed to accompany the guidance for applying the tool that is publicly available in the Cochrane Handbook (Appendix C).¹¹ It should be noted that the ROB

tool has been slightly modified since we started this project in 2010 and new guidelines are available.² In this project we used the original ROB tool. We planned for pilot testing of an additional sample of five trials if there was substantial disagreement. This was not deemed necessary after the initial pilot testing phase.

Table 1. Overview of study components

Study component	Number of reviewers	Number of studies
Assess reliability between individual reviewers applying the Risk of Bias tool	2 reviewers/study 12 reviewers at 2 EPCs	154 RCTs
Assess reliability between consensus agreements of two individual reviewers applying the Risk of Bias tool	2 reviewers/study with consensus 9 reviewers at 4 EPCs	30 RCTs
Assess reliability between individual reviewers applying the Newcastle Ottawa Scale	2 reviewers/study 16 reviewers at 2 EPCs	131 cohort studies

EPC=Evidence-based Practice Center; RCT=randomized controlled trial

Data Extraction

For each trial, the primary outcome was identified and the data necessary to calculate effect estimates were extracted. Several characteristics of the trial that may also be related to risk of bias/quality were extracted, including study type (efficacy, equivalence), study design (parallel, crossover), the condition being treated, nature of the intervention (pharmacological, nonpharmacological), treatment mode (flexible dose vs. fixed dose), treatment duration, type of outcome (subjective, objective), baseline mean difference between study groups for continuous outcomes, the impact of the intervention (treatment ES), variance in ES, sample size, and funding source (Appendix D). This list of variables was compiled prior to commencing data extraction with input from the Steering Committee.

Data extraction for each study was completed at the xx EPC by a single reviewer. A 10 percent random sample of trials with extracted data, including 10 percent of the trials assessed by each reviewer, was checked by a second reviewer. We planned to check an additional 10 percent if there were important or consistent errors, inaccuracies, or omissions. This was not deemed necessary, as there were few errors found.

Data Analysis

Reliability of the ROB tool. For the entire sample of trials, inter-rater agreement between two reviewers was calculated for each domain using weighted kappa statistics. Agreement was categorized as poor, slight, fair, moderate, substantial, or almost perfect using accepted approaches (Table 2).³ The individual kappa statistics for each ROB item are presented and summarized. For the subset of 30 studies, agreement for consensus assessments across pairs of reviewers was assessed using unweighted kappa statistics (i.e., the consensus assessments were compared across the pairs of reviewers from four EPCs).

Table 2. Interpretation of Fleiss' kappa (κ) (from Landis and Koch 1977)³

K	Interpretation
<0	Poor agreement
0.0-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.0	Almost perfect agreement

Validity of the ROB tool. Since there is no gold standard against which the validity of the ROB assessments can be made, the empirically shown inverse association between the ES and the study quality based on ROB assessments was taken as construct validity.

For each RCT we calculated an ES for the primary outcome. ES were calculated using Cohen's d for continuous outcomes. Odds ratios were calculated for dichotomous outcomes and converted into ES using a method developed by Hasselblad and Hedges.⁴ The ES from all RCTs were then combined using a random effects model.⁵ We compared the pooled ES for the high, unclear, and low risk of bias categories for each of the six domains and overall risk of bias. The differences were compared statistically using Kruskal-Wallis and Spearman tests.

The effect of specific covariates on risk of bias was analyzed using a logistic regression. We also tested these covariates for their effect on the association between risk of bias and ES in a subgroup analysis. The covariates examined were intervention type (pharmacological or nonpharmacological), nature of the intervention (behavioral/psychological, device, drug, natural health product, surgical, vaccine, other), study design (parallel vs. other), funding source (industry vs. other), type of trial (efficacy/superiority vs. other), and nature of outcome (subjective or objective).

Newcastle-Ottawa Scale and Cohort Studies

Study Selection

We used an iterative approach to identify a sample of cohort studies based on meta-analyses of cohort studies. Initially, we searched completed EPC reports to identify meta-analyses of cohort studies. We found 3 EPC reports²²⁻²⁴ including 36 cohort studies that met the inclusion criteria (see below). We subsequently conducted searches in Medline using search terms to capture systematic reviews (meta-analys?s.mp, review.pt and search.tw), cohort studies (exp Cohort Studies/, cohort\$.tw, (observation\$ adj stud\$).tw) and meta-analyses (exp meta-analysis/, (analysis adj3 (group\$ or pool\$)).tw, (forest adj plot\$).mp). Results were limited to English language studies in humans that were published in 2000 or later. We searched by year starting with the most recent, and continued until we identified a sufficient number of studies.

A meta-analysis was considered appropriate to include if it had at least 10 cohort studies, assessed a dichotomous outcome, and had substantial statistical heterogeneity (i.e., $I^2 > 50$ percent). Previous meta-epidemiological research has used a minimum sample size per meta-analysis of 5 to 10 studies.^{6,7} This ensures that there is a sufficient pool of studies with some degree of variability in each meta-analysis in order to test the hypotheses. Some degree of heterogeneity is required in order to test whether quality, as assessed by the NOS, can differentiate studies with different effect estimates.

Our target sample size was 125 cohort studies. Initially, 144 cohort studies from 8 meta-analyses were identified; however, 13 studies were not assessed because they were later determined to be the incorrect study design (4 RCT²⁵⁻²⁸; 6 case series/case-controls²⁹⁻³⁴), or they could not be retrieved (3³⁵⁻³⁷). Our final sample included 131 cohort studies (Appendix E).

Quality Assessments

All studies were independently assessed by two reviewers using the NOS. One reviewer was from the xx EPC and one reviewer was from the xx EPC. Discrepancies were resolved through discussion to produce consensus assessments for each study.

Reviewers pilot tested the NOS on three studies³⁸⁻⁴⁰ and met by teleconference to discuss any disagreements in general interpretation of the tool. Decision rules were developed to accompany existing guidance for the NOS (Appendix F and G). A priori we asked clinical experts to provide the minimum length of followup for each review question (Appendix G). We planned for pilot testing of an additional sample of studies if there was substantial disagreement. This was not deemed necessary after the initial pilot testing phase.

Data Extraction

The outcomes and data for effect estimates were based on the meta-analysis and checked against the primary studies by a single reviewer. The statistician double-checked data that were unclear.

Data Analysis

Reliability of the NOS. Inter-rater agreement was calculated for each domain and for overall quality assessment using weighted or unweighted kappa statistics, as appropriate. Agreement was categorized as above.³

Validity of the NOS. For the results of the individual meta-analyses, we coded endpoints consistently so that the outcome occurrence was undesired (e.g., death vs. survival). Within each meta-analysis, we generated a ratio of odds ratio (i.e., odds ratios for studies with and without the domain of interest or of high/low quality as assessed by the NOS). To maintain consistency, we used odds ratios to summarize all meta-analyses, even if this was not the statistic that was used in the original meta-analysis. The ratios of odds ratios for each meta-analysis were combined to give an overall estimate of differences in effect estimates using meta-analytic techniques with inverse-variance weighting and a random effects model.⁸

Results

Overview

The results are presented according to the tools we examined: Risk of Bias (ROB) tool for randomized controlled trials (RCTs) and Newcastle-Ottawa Scale (NOS) for cohort studies. Within each group we present a description of the reviewers involved in performing assessments, a description of the sample of studies assessed, and the results of the reliability and validity analyses, respectively.

Risk of Bias and Randomized Controlled Trials

Description of Reviewers

Twelve reviewers from the xx EPC and the xx EPC assessed the RCTs using the ROB tool. These individuals had varying levels of relevant training, experience with systematic reviews in general, and experience with EPC work specifically. The length of time they had worked with their respective EPC ranged from 9 months to 10 years. Ten of the 12 reviewers had formal training in systematic reviews (i.e., they had taken a university course or attended a Cochrane workshop in systematic reviews). Three of the reviewers had a doctoral degree in epidemiology or health/clinical sciences; eight reviewers had a master's degree in epidemiology/public health, health/clinical sciences, or math/statistics; and one reviewer had an undergraduate degree in health sciences.

For the subset of 30 RCTs, two reviewers from each of the four EPCs were involved in applying the ROB tool and reaching consensus for each study. The reviewers had the following backgrounds: PhD (n=4), MD (n=1), PhD students with completed master's degrees (n=2), MD and master's degree (n=1), and master's degree (n=1). The length of time they had worked with an EPC ranged from 2 to 10 years. Six reviewers had formal training in SRs.

Description of Randomized Controlled Trials

We included 154 RCTs (Appendix B). Details of the trials overall and by risk of bias assessments are provided in Appendix G.

The majority of trials were published in specialty medical journals (87.7 percent). The median impact factor of the journal was 2.9 (interquartile range [IQR] 1.8, 5.1) and the mean number of authors was 7 (standard deviation 3.3). The countries represented most frequently were the United States (31.8 percent), Italy (8.4 percent), and the United Kingdom (8.4 percent). The majority of trials were performed in a single center (74 percent). Approximately 70 percent of trials declared a funding source: industry (27.3 percent) and government (26.0 percent) sources were most frequent.

The design of most trials was parallel (81.8 percent), efficacy/superiority (84.4 percent) with individuals as the unit of randomization (95.5 percent). Just over half of the trials examined drug interventions (53.3 percent), with behavioral/psychological (11.0 percent) and surgical (11.7 percent) interventions commonly represented. Only 35.7 percent of the trials were placebo-controlled. The median sample size was 63 (IQR 39, 123).

A wide range of diagnostic categories was represented (Appendix H). These were classified into: Aging; Cancer Research; Circulatory and Respiratory Health; Gender and Health; Genetics;

Health Services and Policy Research; Human Development, Child and Youth Health; Infection and Immunity; Musculoskeletal Health and Arthritis; Neurosciences, Mental Health, and Addiction; Nutrition, Metabolism, and Diabetes; and Population and Public Health. The most frequently represented categories were Circulatory and Respiratory Health (18.2 percent), Nutrition, Metabolism, and Diabetes (17.5 percent), and Musculoskeletal Health and Arthritis (14.9 percent). The primary outcomes were objective in 48.1 percent of trials and subjective in 51.9 percent. Source of outcome assessment was primarily by clinician (35.1 percent), laboratory measure (23.4 percent), or self-report (23.4 percent).

The vast majority of trials had overall risk of bias assessments of high (46.8 percent) or unclear (52.6 percent) with only one trial assessed as low risk of bias (0.7 percent). Table 3 provides details on the risk of bias assessments for the individual domains. The domains that were most frequently rated as low risk of bias were sequence generation (54.6 percent), missing outcome data (63.6 percent), and selective reporting (77.3 percent). The remaining domains were most frequently assessed as unclear: allocation concealment (77.3 percent), blinding (48.7 percent), and “other sources of bias” (55.8 percent). These results should be interpreted with caution given the low level of agreement between reviewers (Table 4).

We explored study-level variables and their association with domain-specific and overall risk of bias. *Sequence generation* was influenced by the nature of the outcome (objective outcomes showed higher risk of bias, $p=0.01$) and study design (parallel showed lower risk of bias, $p=0.02$). *Allocation concealment* was also influenced by the nature of the outcome with objective outcomes having higher risk of bias ($p=0.0007$). *Blinding* was influenced by nature of the intervention with pharmaceutical interventions having lower risk of bias ($p=0.01$). No variables were associated with risk of bias for incomplete data reporting. *Selective outcome reporting* was associated with the nature of the intervention (surgical trials showed higher risk of bias, $p=0.002$) and funding (industry support had higher risk of bias, $p=0.04$). “*Other sources of bias*” was associated with funding, with industry funding showing higher risk of bias ($p<0.0001$). Finally, *overall risk of bias* was also associated with funding, with industry funding showing higher risk of bias ($p<0.0001$). Of note, “other sources of bias” incorporates several considerations including “inappropriate influence of the study sponsor” (i.e., the extent and nature of involvement of the study sponsor and whether this would likely lead to biased results, Appendix C) which is different from source of funding (i.e., whether the study was funded by industry).

Table 3. Risk of bias assessments by domain* (N=154)

Domain	Risk of bias assessments – n (%)		
	High	Unclear	Low
Sequence generation	0 (0.0)	70 (45.5)	84 (54.6)
Allocation concealment	2 (1.3)	119 (77.3)	33 (21.4)
Blinding	21 (13.6)	75 (48.7)	58 (37.7)
Incomplete data	29 (18.8)	27 (17.5)	98 (63.6)
Selective reporting	16 (10.4)	19 (12.3)	119 (77.3)
Other sources of bias	33 (21.4)	86 (55.8)	35 (22.7)
Overall risk of bias [†]	72 (46.8)	81 (52.6)	1 (0.7)

* The risk of bias assessments presented here are based on consensus between two reviewers.

† Items considered in “overall risk of bias” included design-specific risks of bias; early stopping for benefit; severe baseline imbalances; inappropriate influence of funders (Appendix C).

Inter-rater Reliability

Inter-rater reliability for the RCTs is presented by domain in Table 4. Sequence generation had the highest level of agreement which was considered substantial. Reliability for the remaining domains was considered fair.

Table 4. Inter-rater reliability on risk of bias assessments, by domain (N=154)

Domain	Agreement (weighted κ)	Interpretation ³
Sequence generation	0.79	Substantial
Allocation concealment	0.24	Fair
Blinding	0.33	Fair
Incomplete data	0.34	Fair
Selective reporting	0.27	Fair
Other sources of bias	0.24	Fair
Overall risk of bias	0.26	Fair

We assessed whether important study-level variables influenced inter-rater reliability (Table 5). Assessments for *sequence generation* and *incomplete outcome data* were not influenced by any variable. For *allocation concealment*, inter-rater agreement was better for trials with parallel versus other designs (e.g., crossover, factorial), and for those without versus with industry funding. In terms of *blinding*, inter-rater agreement was better for objective versus subjective outcomes, and trials with other versus parallel designs. For *selective outcome reporting*, inter-rater agreement was greater for trials with hypotheses of efficacy/superiority versus others (e.g., equivalence, non-inferiority). For “*other sources of bias*,” inter-rater agreement was better for trials examining pharmacological versus nonpharmacological interventions and subjective versus objective outcomes.

Table 5. Inter-rater reliability on risk of bias assessments, by domain and study-level variable

Variable	Risk of bias domain, κ (interpretation)*					
	SG	AC	Blinding	Incomplete outcome data	SOR	Other
<i>Overall</i>	0.79 (Su)	0.24 (F)	0.33 (F)	0.34 (F)	0.27 (F)	0.24 (F)
<i>Nature of intervention</i>						
Pharmacological	0.82 (AP)	0.26 (F)	0.33 (F)	0.36 (F)	0.12 (SI)	0.38 (F)
Nonpharmacological	0.77 (Su)	0.24 (F)	0.37 (F)	0.26 (F)	0.36 (F)	-0.06 (P)
p-value	0.57	0.94	0.79	0.61	0.25	0.02
<i>Nature of outcome</i>						
Objective	0.71 (Su)	0.22 (F)	0.54 (M)	0.33 (F)	0.41 (M)	0.09 (SI)
Subjective	0.88 (AP)	0.27 (F)	0.18 (SI)	0.32 (F)	0.07 (SI)	0.45 (M)
p-value	0.09	0.81	0.02	0.93	0.09	0.04
<i>Study design</i>						
Parallel	0.78 (Su)	0.32 (F)	0.27 (F)	0.32 (F)	0.21 (F)	0.23 (F)
Other	0.88 (AP)	-0.07 (P)	0.77 (Su)	0.34 (F)	0.46 (M)	0.30 (F)
p-value	0.47	0.0002	0.0004	0.94	0.26	0.75
<i>Trial hypothesis</i>						
Efficacy/superiority	0.79 (Su)	0.24 (SI)	0.33 (F)	0.32 (F)	0.38 (F)	0.25 (F)
Other	0.81 (AP)	0.29 (F)	0.44 (M)	0.33 (F)	-0.31 (P)	0.31 (F)
p-value	0.92	0.83	0.64	0.97	<0.0001	0.79
<i>Funding</i>						
Industry	0.63 (M)	-0.10 (P)	0.52 (M)	0.42 (M)	0.51 (M)	0.39 (F)
No industry	0.85 (AP)	0.38 (F)	0.28 (F)	0.29 (F)	0.13 (SI)	0.21 (F)
p-value	0.10	0.03	0.21	0.50	0.09	0.34

*AP=almost perfect, Su=substantial, M=moderate, F=fair, Sl=slight, P=poor;
AC = allocation concealment; SG = sequence generation; SOR = selective outcome reporting

Inter-consensus Reliability

A random sample of 30 studies was selected to compare consensus assessments across pairs of reviewers from the four participating EPCs. The results of the inter-EPC reliability are detailed in Table 6. There was moderate agreement for *sequence generation*, fair agreement for *allocation concealment* and “*other sources of bias*,” and slight agreement for the remaining domains and overall risk of bias.

Table 6. Inter-rater reliability between pairs of reviewers on risk of bias assessments across 4 EPCs (N=30)

Domain	Agreement (κ)	Interpretation ³
Sequence generation	0.60	Moderate
Allocation concealment	0.37	Fair
Blinding	0.09	Slight
Incomplete data	0.05	Slight
Selective reporting	0.08	Slight
Other sources of bias	0.27	Fair
Overall risk of bias	0.10	Slight

Validity

Figures 1 to 7 show the effect estimates for studies categorized as high, unclear, and low risk of bias. No statistically significant differences were found in effect sizes (ES) across the risk of bias categories for the six individual domains or overall risk of bias. In five of the seven cases, studies in the high risk of bias category had average ES greater than studies that were low risk of bias. The exceptions were allocation concealment and “other sources of bias.” In six of the seven cases, studies at unclear risk of bias had average ES greater than studies at low risk of bias; in the remaining case (incomplete outcome data) the ES were the same. In four cases, the ES for unclear studies were greater than for high risk of bias (sequence generation, allocation concealment, blinding, “other sources of bias”), and in one case the ES were the same (selective outcome reporting). There was no impact when controlling for study-level factors (i.e., no statistically significant differences were found).

Figure 1. Sequence generation (p=0.50 (Kruskal-Wallis); p=0.21 (Spearman))

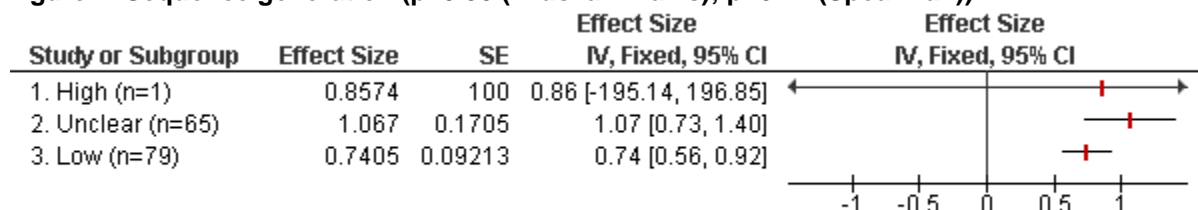


Figure 2. Allocation concealment (p=0.39 (Kruskal-Wallis); p=0.23 (Spearman))

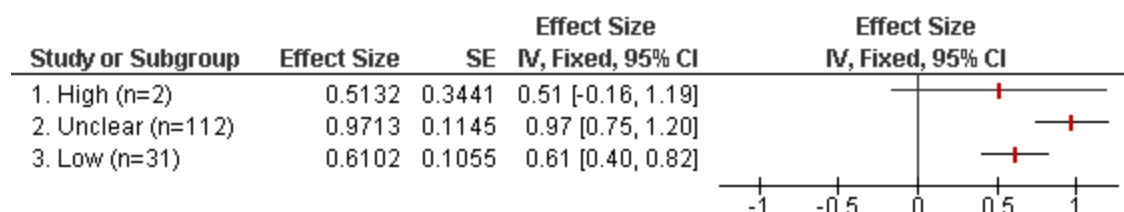


Figure 3. Blinding ($p=0.77$ (Kruskal-Wallis); $p=0.70$ (Spearman))

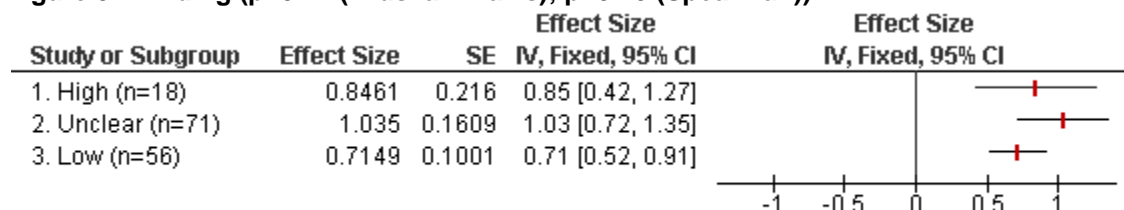


Figure 4. Incomplete outcome data ($p=0.93$ (Kruskal-Wallis); $p=0.99$ (Spearman))

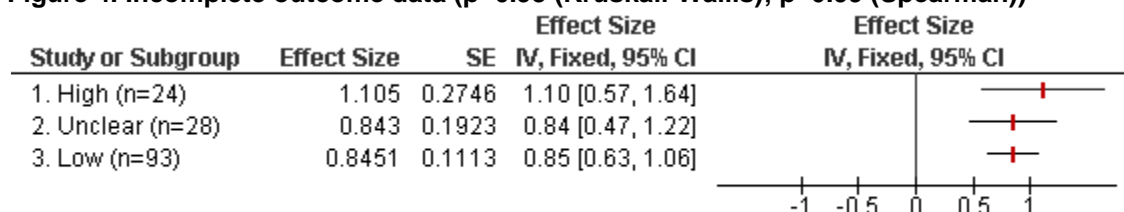


Figure 5. Selective outcome reporting ($p=0.10$ (Kruskal-Wallis); $p=0.12$ (Spearman))

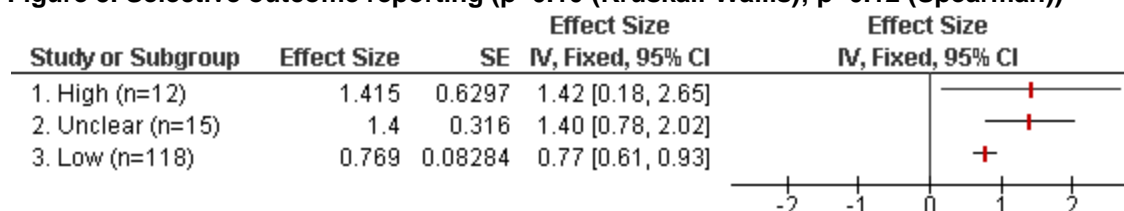


Figure 6. Other sources of bias ($p=0.88$ (Kruskal-Wallis); $p=0.63$ (Spearman))

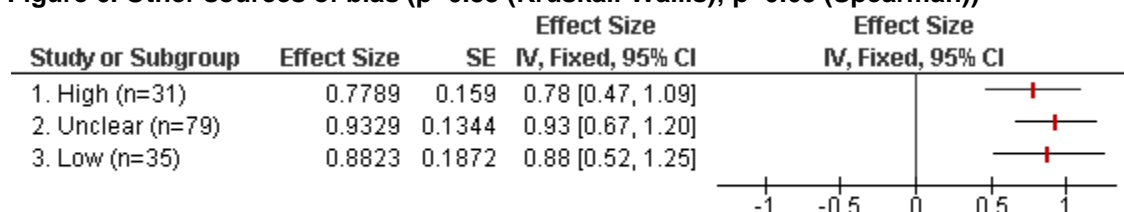
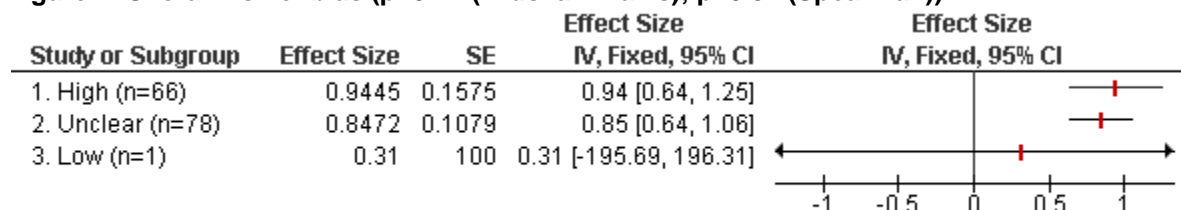


Figure 7. Overall risk of bias ($p=0.71$ (Kruskal-Wallis); $p=0.57$ (Spearman))



Note: The standard errors (SE) for the “high” category in Figure 1 and for the “low” category in Figure 7 were not estimable because there was only one study in each. The SE was made arbitrarily large to reflect the confidence one should have in this estimate.

Newcastle-Ottawa Scale and Cohort Studies

Description of Reviewers

Sixteen reviewers from the xx EPC and the xx EPC assessed the studies using the NOS. Individuals had varying levels of relevant training, experience with systematic reviews in general, and experience with EPC work specifically. The length of time they had worked with their respective EPC ranged from 4 months to 10 years. Thirteen reviewers had formal training in systematic reviews. Four reviewers had a doctoral degree; 10 reviewers had a master’s degree; 1 reviewer had a medical degree and master’s degree; and 1 reviewer had an undergraduate degree.

Description of Sample

The cohort studies were taken from eight meta-analyses which are described in Table 7. Further details are available in Appendix G.

Table 7. Description of meta-analyses of cohort studies included in sample

Topic area (See Appendix G for citations)	Source	Number of studies Included in our sample
Breastfeeding and asthma	EPC report	10
Impaired glucose tolerance and diabetes mellitus	EPC report	17
Cardiac resynchronization therapy and all-cause mortality	EPC report	11
Drug-resistant tuberculosis and positive treatment outcome	Medline	13
Statins and mortality from severe infections and sepsis	Medline	20
Red meat intake and prostate cancer	Medline	15
Overweight and obesity and preterm birth before 37 weeks	Medline	38
Antenatal depression and preterm birth	Medline	20

Inter-rater Reliability

Inter-rater reliability for the 131 cohort studies is presented by domain in Table 8. The item “*was the followup long enough for the outcome to occur*” had the highest level of agreement which was considered substantial. Reliability was moderate for both *ascertainment of exposure* and *ascertainment of outcome*. Reliability was fair for *representativeness of the cohort*, and slight for *comparability of cohorts* and *adequacy of followup of cohorts*. *Selection of the non-exposed cohort* and *demonstration that the outcome was not present at the outset of the study* had poor reliability. Reliability for the overall score (total number of stars) was fair.

Table 8. Inter-rater reliability on NOS assessments, by domain

Domain	Agreement (κ)*	Interpretation ³
Representativeness of the exposed cohort	0.23	Fair
Selection of the non-exposed cohort	-0.03	Poor
Ascertainment of exposure	0.43	Moderate
Demonstration that the outcome was not present at outset of study	-0.06	Poor
Comparability	0.18	Slight
Assessment of outcome	0.49	Moderate

Length of follow-up sufficient	0.68	Substantial
Adequacy of participant followup	0.29	Fair
Total stars	0.29*	Fair

NA=not applicable

* We used a weighted kappa for the total score as it assumes some ordinality in the assessment; other kappas are not weighted, i.e., Cohen's kappa.

Validity

We found no association between individual NOS items or overall NOS score and effect estimates (Table 9).

Table 9. Results of meta-meta-analysis of quality items and measures of association

Domain	ROR	95% CI
Representativeness of the exposed cohort	1.01	0.85, 1.20
Selection of the non-exposed cohort	1.83	0.92, 3.64
Ascertainment of exposure	1.13	0.93, 1.37
Demonstration that the outcome was not present at outset of study	0.72	0.49, 1.07
Comparability	0.86	0.56, 1.31
Assessment of outcome	1.04	0.79, 1.38
Length of followup adequate	0.84	0.55, 1.27
Adequacy of participant followup	0.99	0.91, 1.08

ROR (ratios of odds ratios) that are greater than 1 indicate that studies of higher quality had larger effect sizes on average than studies with lower quality. The RORs presented were pooled across all of the eight meta-analyses that provided data for that quality item; if all studies in a meta-analysis were rated the same for a quality item, that meta-analysis did not contribute to that ROR.

Summary and Discussion

Key Points

Risk of Bias Tool and Randomized Controlled Trials

- Inter-rater reliability between reviewers was fair for all domains except sequence generation which was substantial.
- Inter-rater reliability between pairs of reviewers was moderate for sequence generation, fair for allocation concealment and “other sources of bias,” and slight for the remaining domains.
- Low agreement between reviewers suggests the need for more specific guidance regarding interpretation and application of the Risk of Bias (ROB) tool or possibly re-phrasing of items for clarity.
- Examination of study-level variables and their association with inter-rater agreement identifies areas that require specific guidance in applying the ROB tool. For example, nature of the outcome (objective vs. subjective), study design (parallel vs. other), and trial hypothesis (efficacy/superiority vs. other).
- Low agreement between pairs of reviewers indicates the potential for inconsistent application and interpretation of the ROB tool across different groups and systematic reviews.

- The majority of trials in the sample were assessed as high or unclear risk of bias for many domains. This raises concerns about the methodological rigor of trials in general, and the ability of the ROB tool to detect differences across trials that may relate to biases in estimates of treatment effects.
- No statistically significant differences were found in effect sizes (ES) across high, unclear and low risk of bias categories; however, trends consistently showed greater effect estimates for studies at high or unclear risk of bias.

Newcastle-Ottawa Scale and Cohort Studies

- Inter-rater reliability between reviewers ranged from poor to substantial, but was poor or fair for the majority of domains.
- No association was found between individual quality domains and measures of association.

Discussion

Risk of Bias Tool and Randomized Controlled Trials

We found that inter-rater reliability between reviewers was low for all but one domain in the ROB tool. These findings are similar to results of a previous study⁹ (Table 10). The sample of trials was distinct for the previous and current studies, focusing on pediatric and adult populations, respectively. The common feature of the two samples was that the trials were not part of a systematic review, rather they were trials randomly selected from a larger pool. Hence, the trials covered a wide range of topics. This may have contributed to some of the low agreement as reviewers had to consider different nuances for each trial. Hartling et al. showed improved agreement within the context of a systematic review where all trials examined the same interventions in similar populations¹⁰ (Table 10).

Nevertheless, the low agreement raises concerns and points to the need for clear and detailed guidance in terms of applying the ROB tool. Despite pilot testing and providing supplemental guidance for this study, we still found low agreement. This is likely due to nuances encountered in individual studies. A compilation of examples, especially problem areas, with information on how experts would interpret and apply domains would be of particular benefit for this field. One of the unique contributions of the present study was the analysis of inter-rater reliability controlling for study-level variables. This provides some direction as to where more specific guidance may be beneficial. For instance, agreement was considerably lower for: allocation concealment when trials did not have a parallel design; blinding when the nature of the outcome was subjective; selective outcome reporting when the trial hypothesis was not one of efficacy/superiority; and “other sources of bias” for nonpharmacological interventions and when the outcome was subjective. In summary, agreement for some domains may be better in classic parallel trials of pharmacological interventions, whereas trials with different design features (e.g., crossover) or hypotheses (e.g., equivalence, non-inferiority), and those examining nonpharmacological interventions appear to create more ambiguity for risk of bias assessments.

Another unique contribution of the present study was the examination of the consensus ratings across pairs of reviewers. These ratings should be free of individual rater errors and bias given that these are consensus ratings with disagreements resolved. Further, this is a more meaningful measure of agreement (as opposed to reliability between two reviewers), as these ratings are the ones reported in systematic reviews. In this study, the pairs of reviewers were

from four different centers, each with a long history of producing systematic reviews. The agreement across the pairs of reviewers was generally lower than the agreement between reviewers. This raises concerns about the variability in interpreting and applying the ROB tool that can occur across different systematic review groups and across systematic reviews. It also raises questions regarding the credibility of the risk of bias assessments within any given systematic review.

Table 10. Inter-rater reliability on risk of bias assessments, comparison across studies

Domain	Hartling et al (2009 ⁹)	Hartling et al (2011 ¹⁰)	This study (between reviewers)	This study (between pairs of reviewers)
Sequence generation	Substantial	Almost perfect	Substantial	Moderate
Allocation concealment	Moderate	Moderate	Fair	Fair
Blinding	Fair	Substantial	Fair	Slight
Incomplete data	Fair	Moderate	Fair	Slight
Selective reporting	Slight	Fair	Fair	Slight
Other sources of bias	Fair	Moderate	Fair	Fair
Overall risk of bias	Fair	Moderate	Fair	Slight

Risk of bias for the sample of trials used for this study is described in Table 11 and is compared with samples from other studies. Of particular note is that 99 percent of this sample had overall risk of bias assessments as high or unclear. This is similar to three of the four other samples that had more than 90 percent assessed as high or unclear risk of bias overall (the fourth sample did not assess overall risk of bias). This raises the question of whether all these trials are in fact substantially flawed or whether the ROB tool is overly punitive. If the vast majority of trials are assessed as high or unclear risk of bias, the tool may not be sufficiently sensitive to differences in methodology that might explain variation in treatment effect estimates across studies, or study methodology as a potential explanation for heterogeneity in meta-analyses. Questions also arise regarding whether poor assessments are a result of inadequate or unclear reporting at the trial level. While the focus of the ROB tool is intended to be on methods rather than reporting, reviewers regularly indicate that they rely on the trial reporting to make their assessments. Even within recent samples of trials published after the emergence and widespread dissemination of reporting guidelines,⁴¹ we see large proportions assessed as high or unclear risk of bias. This is consistent with other recent reports of unacceptable reporting in trials.¹ The risk of bias assessments were less severe within the individual domains. However, for the current sample the majority of trials were assessed as high or unclear risk of bias for three of the six domains, including allocation concealment, blinding, and “other sources of bias.” These findings may be beneficial for developers and promoters of reporting guidelines, as well as for researchers who are reporting randomized trials.

Our sampling allowed us to broadly compare our assessments with those of another independent research team.¹ The other team did not apply the risk of bias tool but did assess some of the same domains. Further, the other team examined a larger sample of trials published in 2006 from which our sample was randomly drawn. Nevertheless, the assessment between research teams was consistent for several domains. They found that 75 percent of trials did not report their method of allocation concealment while we found that 79 percent were at high or unclear risk of bias for allocation concealment. Likewise, they found that 59 percent of reports were either not blinded or methods of blinding were not reported while we found that 62 percent of trials were at high or unclear risk of bias for blinding. They found that attrition (intention-to-

treat analysis) was not reported in 31 percent of trials while we found incomplete outcome data for 36 percent. There was variation for one of the domains that both groups assessed: the other team found that sequence generation was not reported for 66 percent of the sample, whereas we found high or unclear risk of bias for sequence generation in only 46 percent of our sub-sample.

Table 11. Trials at high or unclear risk of bias across samples

Domain	Pediatric trials published in late 1990's ⁹ (n=163)	Trials of LABA/ICS in asthma ¹⁰ (n=107)	Pediatric trials published in 2007 ⁴² (n=300)	Pediatric trials published in high impact journals ⁴³ (n=146)	This study (adult trials published in 2006, n=154)
Sequence generation	68%	75%	51%	41%	46%
Allocation concealment	68%	88%	75%	57%	79%
Blinding	40%	58%	50%	19%	62%
Incomplete data	47%	62%	38%	11%	36%
Selective reporting	32%	22%	18%	2%	23%
Other sources of bias	61%	99%	66%	2%	77%
Overall risk of bias	96%	100%	92%	n/a	99%

LABA/ICS = long-acting beta agonists/inhaled corticosteroids

We found no statistically significant association between effect estimates and risk of bias assessments. The main explanations for this finding are that there is in fact no association, or more likely, there was insufficient power to detect differences. One of the factors contributing to low power was the small number of studies within certain domains in the low risk of bias category. This was particularly the case for overall risk of bias as there was only one study in the low category. However, the trend was evident in that the studies at high and unclear risk of bias overall had substantially greater treatment ES (ES=0.94 and 0.85, respectively vs. 0.31). The trend for five of the seven domains (including overall risk of bias) was for greater treatment ES for studies at high risk of bias compared to low risk of bias. Further, in all but one domain, studies at unclear risk of bias had greater treatment ES than studies at low risk of bias, although the differences were not statistically significant. This finding is important in interpreting evidence: when risk of bias is unclear, estimates are likely to be overestimating treatment effects.

Newcastle-Ottawa Scale and Cohort Studies

This is the first study to our knowledge that has examined inter-rater reliability and construct validity of the Newcastle-Ottawa Scale (NOS). We found a wide range of agreement across the domains of the NOS, ranging from slight to substantial. The domain with substantial agreement was not surprising. This domain asked “was the followup long enough for the outcome to occur?” A priori we asked clinical experts to provide the minimum length of followup for each review question. Thus, the assessors had very specific guidance for this item. The agreement for ascertainment of exposure and assessment of outcome was moderate, suggesting that the wording and response options are reasonable. The remaining items had poor, slight, or fair agreement which may be attributable to some of the problems discussed below.

In general, the reviewers found the tool difficult to use. They found the decision rules to be vague, even with the additional information we provided as part of this study. General points that arose were whether to assess each study based on the individual report, or as it related to the systematic review question. For instance, if the systematic review question was specific to a particular population, then the study population may be representative. However, the study

population may not be representative of the average population in the community (first NOS item). Similarly, reviewers wanted specific guidance on whether to base assessments on the information contained in the specific study report, or whether to incorporate information from other reports of the same study. For instance, in numerous cases study authors would refer to another publication for details on the sample or specific methods. Studies could be unnecessarily penalized if they did not incorporate other pertinent information that was available from other reports.

Reviewers found it difficult to determine the difference between some of the response options. For example, two of the response options for item 1 regarding the exposed cohort are “truly” versus “somewhat” representative. Some reviewers questioned whether this distinction was important, as both responses garnered a star for that item, hence there was no difference in the final score. Also with respect to the first item, reviewers were uncertain regarding what makes a population “selected.” Some interpreted this to include populations with unequal representation of a certain group (e.g., 90 percent males, all patients had organ transplant) while others relied on the methods of selection (e.g., volunteers, select group such as nurses). Likewise, reviewers questioned the difference between the categories “structured interview” and “written self-report” for ascertainment of exposure. For example, researchers may use structured, validated surveys or questionnaires (e.g., SF-36) but these are completed independently by the study participant.

Reviewers were uncertain on how to assess the item on comparability. Some studies discussed testing different confounders in their models, but only included the confounders that showed a significant difference in the final model. Reviewers were unsure whether to indicate that the study controlled for that confounder.

Reviewers questioned what some domains actually measured. For instance, whether the selection domain assesses bias in how the participants were selected, or whether it is intended to assess the applicability of the study population to the population in general. Further, some concerns were raised that the response categories within a domain measured different constructs.

Reviewers would have liked “unclear” or “no description” options for some items, in particular for the last item on “adequacy of followup of cohorts.” They identified an additional problem with the response categories for this item. The second option is either a small number lost or description provided of those lost. The third option is a larger number lost and no description of those lost. However, there is no response option that includes a larger number lost *and* a description is provided (e.g., that indicates there was no imbalance between groups).

Finally we found no association between NOS items and the measures of association using meta-epidemiological methods that control for heterogeneity due to condition and intervention. Moreover, we saw no trends suggesting an association between magnitude of association and quality. This provides empirical evidence to substantiate previous claims that “the NOS includes problematic items with an uncertain validity.”⁴⁴

Implications for Practice

The findings of this research have critical implications for practice and the interpretation of evidence. The low level of agreement between reviewers and pairs of reviewers puts into question the credibility or validity of risk of bias/quality assessments using the ROB tool or NOS within any given systematic review. Moreover, in measurement theory, reliability is a necessary condition for validity (i.e., without being reliable a test cannot be valid). Systematic reviewers are urged to incorporate considerations of risk of bias/quality into their results. Furthermore,

integration of the GRADE tool into systematic reviews necessitates the consideration of risk of bias/quality assessments in rating the strength of evidence and ultimately recommendations for practice.⁴⁵ While the ROB tool considers risk of bias for an individual study, the GRADE tool assesses the risk of bias across all relevant studies for a given outcome (e.g., most information is from studies at high/moderate/low risk of bias).⁴⁶ The results of risk of bias assessments and their interpretation in a systematic review, as well as the strength of evidence assessments, will be misleading if they are based on flawed assessments of risk of bias/quality. Moreover, Stang declared with respect to the NOS that “use of this score in evidence-based reviews and meta-analyses may produce highly arbitrary results.”⁴⁴

Reviewers and review teams need to be aware of the limitations of existing tools. Detailed guidelines, decision rules, and transparency are needed so that readers and end-users of systematic reviews can see how the tools were applied. Further, pilot testing and development of review-specific guidelines and decision rules should be mandatory and reported in detail.

This study provides some evidence of association (or trends) between risk of bias domains and estimates of treatment effect which corroborates previous findings.^{9,42} The results confirm that the ROB tool is doing what it is intended to do, i.e., identifying studies that may yield less reliable estimates of treatment effects. We did not find similar evidence for the NOS, therefore its suitability for use in systematic reviews should be re-examined.¹⁷ Further, the NOS in its current form does not appear to provide reliable quality assessments and requires further development and more detailed guidance. The NOS was previously endorsed by The Cochrane Collaboration; however, more recently the Collaboration has proposed a modified ROB tool to be used for nonrandomized studies.² A new tool developed through the EPC Program for quality assessment of nonrandomized studies offers another alternative.⁴⁷

Future Directions

There is a dire need for more detailed guidelines to apply both the ROB tool and the NOS, as well as revisions to the tools to enhance clarity. We have identified specific trial features for which clearer guidance is needed. A living database that collects examples of risk of bias/quality assessments and consensus from a group of experts would be a valuable contribution to this field. Individual review teams and research groups should be encouraged to begin identifying examples and these could be compiled across programs (e.g., the EPC Program) and entities (e.g., The Cochrane Bias Methods Group), and made widely accessible. We have identified specific problems with application and interpretation of the NOS tool. Further revisions and guidance are needed to support the continued use of NOS in systematic reviews. Investment in further reliability and validity testing of other tools may be more appropriate (e.g., Cochrane ROB tool for nonrandomized studies, the EPC tool). Finally, consensus in this field is needed in terms of the threshold for inter-rater reliability of a measurement before it can be used for any purpose, even descriptive purposes (i.e., describing the risk of bias or quality of a set of studies).

Strengths and Limitations

This is one of few studies examining the reliability and validity of the ROB tool. It is the first to our knowledge that examines reliability between the consensus assessments of pairs of reviewers. Further, it is the first study to provide empirical evidence on study-level variables that may impact reliability of ROB assessments. This is the first study to our knowledge that examined reliability and validity of the NOS.

The main limitation of the research is that the sample sizes (154 RCTs, 131 cohort studies) may not have provided sufficient power to detect statistically significant differences in ES estimates according to risk of bias/quality. We observed trends for RCTs, with larger effect estimates for studies at high or unclear versus low risk of bias. We found no significant associations between quality and measures of association within the cohort studies, which could be attributable to low power. However, we did not find any discernable trends. We specifically selected meta-analyses with substantial heterogeneity in order to optimize our potential to see whether quality as assessed with the NOS might explain variations in measures of association.

We involved a number of reviewers with different levels of training, type of training, and extent of experience in quality assessment and systematic reviews. Some of the variability or low agreement may be attributable to characteristics of the reviewers. Agreement may be higher among individuals with more direct experience or specific post-graduate training in research methods or epidemiology. Nevertheless, all reviewers had previous experience in systematic reviews and quality assessments, and likely represent the range of individuals that would typically be involved in these activities within a systematic review.

A final caveat to note is that the ROB tool has undergone some revisions since we initiated the study. These are detailed in the most recent version of the Cochrane Handbook² but were not incorporated into our research. The changes affected primarily the blinding and the “other sources of bias” domains. This does not impact the general findings from our research; however, further testing with the modified tool is warranted.

Conclusions

More specific guidance is needed to apply and interpret risk of bias/quality tools. We identified a number of study-level factors that influence agreement. This information provides direction for more detailed guidance. Low agreement across pairs of reviewers raises questions about the credibility of risk of bias assessments in any given systematic review. This has implications for incorporation of risk of bias into results and grading the strength of evidence. There was variable agreement across items in the NOS. This finding, combined with a lack of evidence that it discriminates studies that may provide biased results, challenges its suitability for use in systematic reviews.

References

1. Hopewell S, Dutton S, Yu LM, et al. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. *BMJ* 2010;340:c723.
2. Cochrane Handbook for Systematic Reviews of Interventions. Higgins PT, Green S, editors. 2011;(5.1.0. [updated March 2011]).
3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-74.
4. Hasselblad V, Hedges LV. Meta-analysis of screening and diagnostic tests. *Psychol Bull* 1995;117(1):167-78.
5. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;7(3):177-88.
6. Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technol Assess* 2003;7(1):1-76.
7. Furukawa TA, Watanabe N, Omori IM, et al. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. *JAMA* 2007;297(5):468-70.
8. Sterne JA, Juni P, Schulz KF, et al. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21(11):1513-24.
9. Hartling L, Ospina M, Liang Y, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. *BMJ* 2009;339:b4012.
10. Hartling L, Bond K, Vandermeer B, et al. Applying the risk of bias tool in a systematic review of combination long-acting beta-agonists and inhaled corticosteroids for persistent asthma. *PLoS One* 2011;6(2):e17242.
11. Cochrane Handbook for Systematic Reviews of Interventions. Higgins PT, Green S, editors. 2009;5.0.2 [updated September 2009].
12. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med* 2010;7(9):e1000326.
13. Moher D, Jadad AR, Nichol G, et al. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. *Control Clin Trials* 1995;16(1):62-73.
14. Juni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999;282(11):1054-60.
15. West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. *Evid Rep Technol Assess (Summ)* 2002;(47):1-11.
16. Olivo SA, Macedo LG, Gadotti IC, et al. Scales to assess the quality of randomized controlled trials: a systematic review. *Phys Ther* 2008;88(2):156-75.
17. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. *Health Technol Assess* 2003;7(27):iii-173.
18. Wells G, Shea B, O'Connell J, Robertson J, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analysis [Website]. http://www.orhi.ca/programs/clinical_epidemiology/oxford.asp. 2011. Available from: URL: http://www.orhi.ca/programs/clinical_epidemiology/oxford.asp.
19. Wells G, Shea B, O'Connell J, Robertson J, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analysis [Abstract]. 2000.

20. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998;52(6):377-84.
21. Wells G, Brodsky L, O'Connell D, Robertson J, et al. Evaluation of the Newcastle-Ottawa Scale (NOS): an assessment tool for evaluating the quality of non-randomized studies[Abstract]. 2003.
22. Ip S, Chung M, Raman G, et al. Breastfeeding and maternal and infant health outcomes in developed countries. *Evid Rep Technol Assess (Full Rep)* 2007;(153):1-186.
23. McAlister FA, Ezekowitz J, Dryden DM, et al. Cardiac resynchronization therapy and implantable cardiac defibrillators in left ventricular systolic dysfunction. *Evid Rep Technol Assess (Full Rep)* 2007;(152):1-199.
24. Santaguida PL, Balion C, Hunt D, et al. Diagnosis, prognosis, and treatment of impaired glucose tolerance and impaired fasting glucose. *Evid Rep Technol Assess (Summ)* 2005;(128):1-11.
25. Buxton AE, Lee KL, Fisher JD, et al. A randomized study of the prevention of sudden death in patients with coronary artery disease. Multicenter Unsustained Tachycardia Trial Investigators. *N Engl J Med* 1999;341(25):1882-90.
26. Sanchez JM, Katsiyannis WT, Gage BF, et al. Implantable cardioverter-defibrillator therapy improves long-term survival in patients with unexplained syncope, cardiomyopathy, and a negative electrophysiologic study. *Heart Rhythm* 2005;2(4):367-73.
27. Frost FJ, Petersen H, Tollestrup K, et al. Influenza and COPD mortality protection as pleiotropic, dose-dependent effects of statins. *Chest* 2007;131(4):1006-12.
28. Tseng MY, Hutchinson PJ, Czosnyka M, et al. Effects of acute pravastatin treatment on intensity of rescue therapy, length of inpatient stay, and 6-month outcome in patients after aneurysmal subarachnoid hemorrhage. *Stroke* 2007;38(5):1545-50.
29. Wisner KL, Sit DK, Hanusa BH, et al. Major depression and antidepressant treatment: impact on pregnancy and neonatal outcomes. *Am J Psychiatry* 2009;166(5):557-66.
30. Suri R, Altshuler L, Hellemann G, et al. Effects of antenatal depression and antidepressant treatment on gestational age at birth and risk of preterm birth. *Am J Psychiatry* 2007;164(8):1206-13.
31. Ancel P, Saurel-Cubizolles M, Di Renzo G, Papiernik E, Breart G. Very and moderate preterm births: are the risk factors different? *Br J Obstet Gynaecol* 1999;106, 1162-70.
32. Schlienger R, Fedson D, Jick S, Jick H, Meier C. Statins and the risk of pneumonia: a population-based, nested case-control study. *Pharmacotherapy* 2007;27, 325-32.
33. Blaas S, Mutterlein R, Weig J, et al. Extensively drug resistant tuberculosis in a high income country: a report of four unrelated cases. *BMC Infect Dis* 2008;8, 60-7.
34. Condos R, Hadgiangelis N, Leibert E, et al. Case series report of a linezolid-containing regimen for extensively drug-resistant tuberculosis. *Chest* 2008;134, 187-92.
35. Rahaman J, Narayansingh G, Roopnarinesingh S. Fetal outcome among obese parturients. *Int J Gynaecol Obstet* 1990;31, 227-30.
36. Dayan J, Creveuil C, Herlicoviez M, et al. [Antenatal depression, a risk factor for prenatal delivery]. *Presse Med* 1999;28(31):1698.
37. Kim Y, Lee B, Park H. Risk factors for preterm birth in Korea: a multicenter prospective study. *Gynecol Obstet Invest* 2005;60, 206-12.
38. Norman R, Masters L, Milner C, Wang J, Davies M. Relative risk of conversion from normoglycaemia to impaired glucose tolerance or non-insulin dependent diabetes mellitus in polycystic ovarian syndrome. *Hum Reprod* 2001;16(9), 1995-8.

39. Wright A, Holberg C, Taussig L, Martinez F. Factors influencing the relation of infant feeding to asthma and recurrent wheeze in childhood. *Thorax* 2001;56(3), 192-7.
40. Arvanitakis Z, Schneider JA, Wilson RS, et al. Statins, incident Alzheimer disease, change in cognitive function, and neuropathology. *Neurology* 2008;70(19 Pt 2):1795-802.
41. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c332.
42. Hamm MP, Hartling L, Milne A, et al. A descriptive analysis of a representative sample of pediatric randomized controlled trials published in 2007. *BMC Pediatr* 2010;10:96.
43. Crocetti MT, Amin DD, Scherer R. Assessment of risk of bias among pediatric randomized controlled trials. *Pediatrics* 2010;126(2):298-305.
44. Stang A. Critical evaluation of the Newcastle-Ottawa scale for the assessment of the quality of nonrandomized studies in meta-analyses. *Eur J Epidemiol* 2010;25(9):603-5.
45. Schunemann HJ, Brozek J, Oxman AD. GRADE handbook for grading quality of evidence and strength of recommendation. Version 3.2 [update March 2009] ed. The Grade Working Group; 2009.
46. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams JW, Atkins D, Meerpohl J, Schunemann HJ. GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias). *J Clin Epidemiol* 2011; 64:407-415.
47. Berkman ND, Viswanathan M. Development of a tool to evaluate the quality of non-randomized studies of interventions or exposures. 2009.

Abbreviations

Abbreviations	Full Text
AHRQ	Agency for Healthcare Research and Quality
CERs	Comparative Effectiveness Reviews
CHIP	Children's Health Insurance Program
EHC	Effective Health Care
EPC	Evidence-based Practice Center
ES	Effect Size(s)
I^2	I-squared
ICC	Intra-Class Correlation
IQR	Inter-Quartile Range
NOS	Newcastle Ottawa Scale
RCTs	Randomized Controlled Trials
ROB	Risk of Bias
ROR	Ratio of Odds Ratios
SR	Systematic Review